Mining Lifestyle Personas at Scale in E-commerce

Kang Li*, Vinay Deolalikar†, and Neeraj Pradhan‡
Search and Data Mining
Groupon
Palo Alto, CA 94306
E-mail: *kli@groupon.com, †vdeolalikar@groupon.com, ‡neepradhan@groupon.com

***Abstract*—Groupon is a major e-commerce company. It is unique in the sense that it is not only a vendor of goods, but also of local deals (such as restaurants, spas, activities, etc.) that reflect various aspects of a user's interests. In this sense, Groupon has a complete view of its users' lifestyle preferences. This is different from e-commerce goods vendors, who, for instance, may not have direct insight into what restaurants are preferred by their users, or what nightlife they prefer. We may say that Groupon engages the entire "lifestyle persona" of its users.**

Motivated by this, we consider the large scale problem of mining such "lifestyle personas" from Groupon's activity data collected over 38 million e-commerce users. This includes users drawn from one of the world's largest mobile user bases. Our solution combines domain knowledge from e-commerce with data mining and graph theoretic methods.

Since Groupon offers deals and goods across the 360 degree gamut of user preferences, do lifestyle personas in its data also span this gamut? What are some of the significant personas that emerge from this data mining? Are the differences between personas gross or subtle? These are some of the questions we answer conclusively in our study.

Our mined personas are both descriptive and distinctive, and offer insights into customer behaviors and lifestyles. Our work is being used to redesign the user experience, as well as to power product recommendations at Groupon.

I. Introduction

Groupon is a major multi-billion dollar e-commerce company. Unlike some e-commerce companies such as Amazon or eBay, Groupon is also a major seller of local deals—-restaurants, cafes, spas, manicures, hair salons, beauty treatments, and so on. In addition, it is a major player in goods e-commerce. Groupon's estimated gross billings for the year 2015 are over USD 6 Billion. Over 50% of those come from local deals, and over 30% from goods.

As background, *local* deals are those that are viewed, clicked and purchased online, and redeemed offline with location constraints, such as deals on spas, restaurants, etc. In contrast, *goods* are viewed, clicked, purchased online, and there is no further redemption required. Examples are purchases of cameras or phones from online sellers. Local and goods are two core components that reflect customers choices: both offline and online. Furthermore, each component appeals to a different aspect of a user's persona.

The presence in both local and goods gives Groupon a unique "360°view" of its customers' *lifestyles.* Not only does Groupon know that a customer is interested in consumer electronics, but also that they take tennis lessons on weekdays, go to a spa on weekends, buy tickets for their children at the local museum events, and so on. Due to its unique position as a leader in local commerce as well as a significant vendor in goods e-commerce, Groupon potentially has insight into the entire personality type of the most engaged (i.e., those users who spend a lot of time on Groupon) mass of its users.

This brings us to our problem statement:

> *Can we mine "lifestyle personas"—personas that capture a significant gamut of user lifestyles—from the aggregate 360°view data at Groupon?*

There is scant literature on this problem statement.[1] There is literature on creating customer segments, on personalized recommendations based on segments, but not (to our knowledge) on the extraction of lifestyle personas from segments. In this paper, we address this gap by answering the following research questions.

1) How do we mine lifestyle personas from a large real-world corpus of tens of millions of users?
2) What do lifestyle personas mined from 360°views of user preferences look like?
3) How distinctive are personas: are the differences between personas stark or subtle?

The hurdles to be overcome in answering these questions are manifold. First, there is the question of scale: our study features 38 million users, making it larger than published customer segmentation studies. Next comes the task of building user representations given multiple domain-specific obstacles such as sparsity of user activity on many categories, time-varying nature of user preferences, capturing the user's interaction with the UI, etc. We share the domain knowledge required to address each of these obstacles. Next, the core task of extracting personas requires the use of data mining techniques such as clustering and frequent itemsets, but in non-standard ways. Parameter choices that affect the user segmentation have to be made (such as, the optimal number of segments). Furthermore, the "denoising" of personas requires graph theoretic tools to be crafted. Finally, the "goodness of fit" of personas is measured. The above also serves as an outline of the main steps in our approach.

Our contributions, broadly speaking, are:

1) We present the first large scale (over 38 million users) segmentation study at a major e-commerce company.

[1]Most e-commerce companies do not have 360°view data, and companies are reluctant to reveal any properties of their customer data to the public.

Although several companies do such studies internally, there are no published studies in this field, depriving the research community of valuable benchmarks. We provide our techniques in detail, and share illustrative examples of mined personas.

2) Our study is the first, in published literature, to associate lifestyle personas to segments.
3) We fashion tools using datamining and graph-theoretic concepts that enable us to tame the many hurdles—small and large—in the mining of lifestyle personas at big data scales.

II. Related Work

In e-commerce, understanding user interests in order to use them for personalized recommendations has received much attention over the past decade. For instance, Amazon.com identifies user interests using item-item collaborative filtering [7] and shows that customers like to buy other items that are similar to their favorite items. [13] mine user interests by clustering user click-stream data. They show that users' visiting sequences, frequencies, and time spent on each category are related to their interests. [17] tackles the problem of learning user interests by applying association rule mining and classification. In [16], researchers expand user interests mining to customer satisfaction mining, using soft segmentation of users. Recently, [9] addressed the problem of user interests mining for new users: they co-cluster demographic data of customers. [3] discusses big-data technologies for learning user interests at scale.

Another related area to ours is user modeling, which seeks to characterize user behaviors. [8] tackle the problem using an ontology-based user model for e-commerce customer demographic data, browsing histories, etc. [11] discuss the software life-cycle aspect of e-commerce user modeling using clustering.

[4, 10] discuss a rather different concept, also called personas. These are descriptions of fictional users that are used for participation in interactive design. Such fictional personas are created using information about users' needs, behaviors, and preferences. In [1], thematic analysis on qualitative data, such as observations and interviews, is used in to create such personas for interaction design. In [14], quantitative data, along with clustering techniques and principal components analysis [5], is used to tackle the problem. Although these studies are superficially related to our work, there are significant differences. Specifically, these studies focus on extracting *fictional* representations of general user profiles in an interactive setting so that designers could then interact with this fictitious persona. In contrast, our work is focused on e-commerce, and it seeks to learn shared interests of groups of users.

Broadly speaking, our work significantly differs from existing work in both the data and the application. First the data: due to the availability of Local and Goods data in our analysis, our user lifestyle personas cover "360°" of user preferences; in contrast existing work in e-commerce, primarily considers Goods preferences only. Second the application: our work seeks to mine high level and human-interpretable personas in e-commerce. Although such personas might implicitly exist in collaborative filtering frameworks in e-commerce, to our knowledge there has not been specific attention given to mining them at real-world scales. That is the gap in literature that our work addresses.

III. Building User Representations

In e-commerce, products are categorized into a product taxonomy (see Fig. 1). Let the set of categories in the product taxonomy be $\{C_1, \ldots, C_M\}$. In this section, we describe four types of adjustments required in order to associate each user to a single M-dimensional activity vector over the set of categories.

A. Time Decays

User interests and behaviors in e-commerce may change across time [12, 2]. Accordingly, we wish to stress recent activity (clicks, purchases, etc.) in the mining of user personas, and de-stress past activity. Therefore, each activity is decayed with a daily time decay factor $\alpha \in (0,1]$. If $A(t)$ is the activity performed by a user at time t, and $t_p (> t)$ is the time when user personas are learned, the the time decayed count of activity $A(t_p)$ is $A(t_p) = A(t)\ \alpha^{t_p - t}$.

B. Position Correction

In e-commerce websites and mobile applications, products that appear on higher positions usually gain significantly more attention from users than products on lower positions. As a result, most clicks and purchases are performed on higher positions. Intuitively, when a user clicks or purchases a deal located at a low position, it indicates a strong interests in the deal. In order to account for this, *position correction* is applied.

We assign positions numbers $P_0, P_1, ..., P_N$ for product places starting from the top to the bottom (of, say, our e-commerce website or mobile application). We process each activity with a position correction boosting factor β applied. The value of β varies with the position at which the product is displayed. This is done in order to capture users' stronger indication of interests when they click or purchase products at lower positions, relative to those at higher positions.

In our system, we have set the following position-adjustment boost values for β:

(a) 1, which means no boosting, for products at position either P_0 or P_1; (b) 3, which means triple counting each activity, for products in position P_2 to P_{20}; (c) 10 for products whose position is lower than P_{20}.

These settings are tuned using the results of A/B tests in our e-commerce system.

C. Activity Integration

For the sake of simplicity, to capture customer preferences, we combine user clicks and purchases: since purchases are a

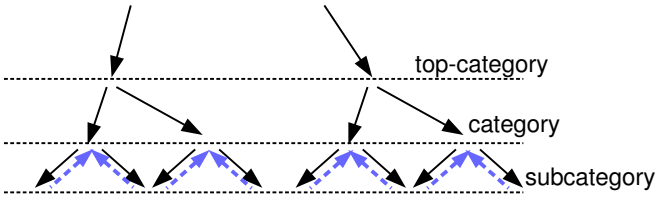

Fig. 1: The DAG of the product taxonomy, showing activity aggregation (in blue arrows) from subcategories to categories. See §III-D.

stronger signal than clicks, we convert purchases into "pseudo-clicks" by converting each purchase into γ (where $\gamma = 5$) clicks. Finally, we add pseudo-clicks with actual clicks to obtain *unified clicks*.

D. Activity Aggregation along Taxonomy

The e-commerce inventory is categorized using a product taxonomy. This taxonomy is a DAG: a top-category has children categories, each of whom further has children subcategories, as shown in Fig. 1.

To reduce sparsity caused by the large number of products, we further aggregate counts of activities on subcategories, so that we have activities over categories. For instance, a user who has 5 clicks on *Motorola Droid Ultra 4G Android Smartphone* and 6 clicks on *Apple iPhone 5*, has 11 clicks on the category *Electronics*. This is because both deals are in the same taxonomy branch described by Top-Category: *Goods* $\rightarrow$ Category: *Electronics* $\rightarrow$ Subcategory: *Cell Phones and Accessories*.

Terminology. For the rest of this paper, *activity* stands for time decayed, position corrected and aggregated *unified click* on categories. We also use the term *persona* in short for *lifestyle persona*.

IV. Mining Personas

There are three stages in extracting personas (Fig. 2): *User Selection and Segmentation*, *Extracting Elements of Personas from Segments*, and *Persona Generation from Elements*. Each stage comprises multiple steps. We now explain each stage conceptually along with examples at each stage.

Implementation details are presented in §VI-A.

A. User Selection and Segmentation

a) Step 1: Select Highly Engaged Users: A pervasive state of affairs in e-commerce is that a significant body of registered users are inactive: they have few clicks and no recent purchases. This includes seasonal and occasional buyers who only use the services in particular time of a year, new users, etc. Including such users in persona mining adds noise. Therefore, users with insufficient activities are filtered from our study of user personas.

We wish to select highly engaged users. Therefore, we consider user activities as the indicator in order to decide whether a user is engaged or not. Specifically, if a user has either purchased any product or has more activity than a pre-defined threshold, the user will be regarded as a highly engaged user, and their activities will be used for persona mining.

Analyzing data from July 1st, 2014 to January 1st, 2015 indicates that 26% of users have purchased at least one product over this six month period. For the remaining 74% of users, the pre-defined activity threshold is applied.

There are two aspects to be considered in selecting the activity threshold: the lower the activity threshold, the more users will be included in the persona mining process; whereas the higher the predefined threshold is, the more activities each selected user has. We wish to set the pre-defined threshold properly, so that enough users are selected, and each selected user has sufficient activities for persona mining.

In our study, we set an activity threshold such that around 25% of users have more clicks than the threshold. After initialization, the threshold is tuned further by means of A/B testing.

b) Step 2: Segment Users using Clustering: Users of different preferences and interests are grouped separately and handled independently. To achieve this goal, users are clustered into K *segments* $\{S_i : 1 \le i \le K\}$. All the remaining steps of the algorithm are conducted on a per-segment basis.

B. Element Extraction from Segments

c) Step 3: Extract Frequent Categorysets: Frequent Categorysets are the adaptation of frequent itemsets from transactional data to our setting.

Definition 1. Consider the segment S, and a set of categories $\mathscr{C}$. The *support* $\text{supp}(\mathscr{C})$ is the set of all users in S that have activity in each $C \in \mathscr{C}$. $\mathscr{C}$ is a *frequent categoryset* with respect to the threshold T if $|\text{supp}(\mathscr{C})|/|S| > T$. Namely, at least T proportion of the users in S have activity in each $C \in \mathscr{C}$.

At this step, we extract frequent categorysets from each segment.

d) Step 4: Retain only Maximal Frequent Categorysets:

Definition 2. A *maximal frequent categoryset* for S is a frequent categoryset that is maximal w.r.t. set inclusion. Namely, it is not a subset of any other frequent categoryset of S.

At this step, within each segment, only maximal frequent categorysets are retained from the the set of frequent categorysets extracted at previous step. For example, if $\{$*iPhone covers, batteries*$\}$ and $\{$*iPhone covers, batteries, iPad accessories*$\}$ are both frequent categorysets for S, then only the latter is retained, since the former is contained in the latter.

e) Step 5: Associate Maximal Frequent Categorysets with Elements of Personas: Intuitively, frequent categorysets represent combinations of products or deals that large groups of users in the segment prefer. These preferences fit well our concept of user lifestyle personas. Therefore *maximal frequent categorysets from a user segment are used as elements of the persona of that segment*. We denote the set of all persona elements from a segment S by $\mathsf{ELEM}(S)$.

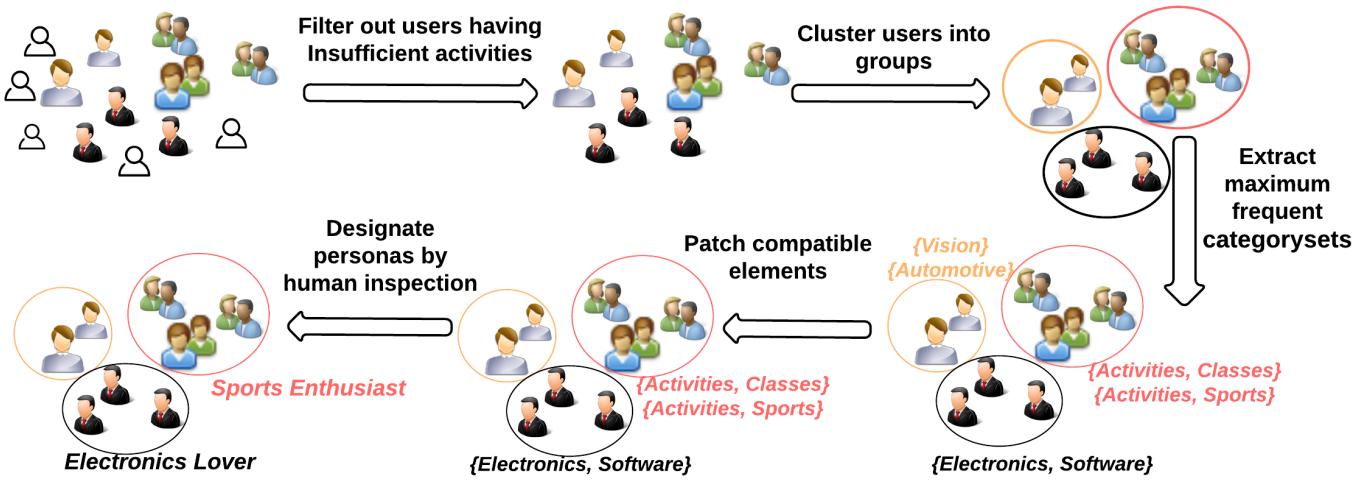

Fig. 2: Major steps in the mining of personas. Users are segmented, elements of personas extracted, and only compatible elements are patched into lifestyle personas. Details in §IV.

C. Generate Personas

f) Step 6: Check if Elements are Compatible using Graph Connectedness: Now, we associate each user segment to (at most) a single persona. We inspect the elements of the persona extracted from the segment in the previous step. We then "patch together" compatible elements to give a persona, as follows (see Fig. 3).

Consider a segment S from which $\mathsf{ELEM}(S)$ has been extracted.

Definition 3. Two elements $p_r, p_s \in \mathsf{ELEM}(S)$ are said to be *compatible* iff $p_r \cap p_s \neq \emptyset$. Namely, if they contain at least one category in common.

To S, we now associate a graph $G(S)$. The vertex set of $G(S)$ is $\mathsf{ELEM}(S)$. It remains to describe the edge set of $G(S)$.

Definition 4. Two vertices (or elements) $p_r, p_s \in \mathsf{ELEM}(S)$ are joined by an edge in $G(S)$ iff they are compatible.

Finally, we transfer the notion of compatibility, via $G(S)$, onto $\mathsf{ELEM}(S)$.

Definition 5. $\mathsf{ELEM}(S)$ is said to be *compatible* iff $G(S)$ is connected.

Note that up till Def. 5, we had discussed only pairwise notions of compatibility. Def. 5 transfers these notions to that of compatibility a set using graph connectivity.

To better explain how to decide if element sets are compatible, we show two examples in Fig. 3 and Fig. 4. The first example discusses a compatible set, whereas the second example discusses an incompatible set, respectively.

In Fig. 3, $\{clothing, jewelry\}$, $\{jewelry, nightlife\}$, and $\{nightlife, bars\}$ is the set of elements (maximal frequent categorysets) for a group of customers. As per Def. 3, the first two elements $\{clothing, jewelry\}$ and $\{jewelry, nightlife\}$ are compatible, because they have one common category $\{jewelry\}$. Similarly, the last two elements $\{jewelry, nightlife\}$ and $\{nightlife, bars\}$ are compatible, because they have one common category $\{nightlife\}$. We represent each element as a node in a graph. As per Def. 4, we join the nodes corresponding to $\{clothing, jewelry\}$ and $\{jewelry, nightlife\}$ with an edge since these two elements are compatible. Similarly, we also join the nodes corresponding to $\{jewelry, nightlife\}$ and $\{nightlife, bars\}$ with an edge. Finally, as per Def. 5, the set $\mathsf{ELEM}(S)$ itself is said to be compatible, because the formed graph is connected as illustrated in Fig. 3.

In Fig. 4, we show an example of an incompatible element set. In this example, $\{outdoors, gyms\}$, $\{gyms, sports\}$, and $\{weight-loss, medical\}$ are the maximal frequent categorysets for another group of customers. By Def. 3, the first two elements $\{outdoors, gyms\}$ and $\{gyms, sports\}$ are compatible, because they have one common category $\{gyms\}$. We then connect the nodes for these two elements in the graph representation. Since there is no edge connecting the third element $\{weight-loss, medical\}$, the formed graph is not connected. By Def. 5, $\mathsf{ELEM}(S)$ is said to be incompatible.

g) Step 7: Decide if Segment has a Persona: If $\mathsf{ELEM}(S)$ are compatible, then they are passed on further to human analysts to suggest a lifestyle persona. In case $\mathsf{ELEM}(S)$ is not compatible, then we simply flag that segment as not having provided us with a coherent signal to associate a persona to it. The reason for being conservative in this regard is that personas are used primarily for targeting users with recommendations, discounts, etc. Therefore, it is advisable not to incorrectly associate a significant number of users with persona. In practice however, we found only few clusters that do not admit a persona.

Returning to our running examples, the set of maximal frequent categorysets in Fig. 3 is said to be compatible, therefore, it will be passed to human analysts; while the set of maximal frequent categorysets in Fig. 4 is said to be

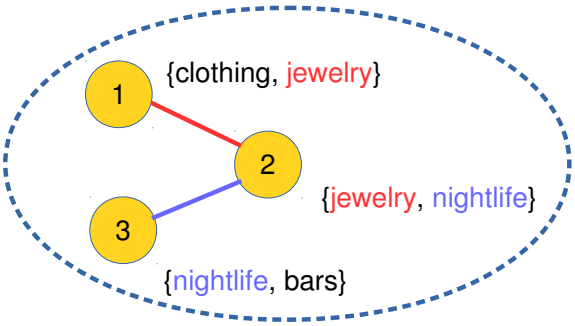

Fig. 3: Patching of compatible elements to form a persona (§IV-C). Each element is the vertex of a graph. Vertices are joined if the elements overlap. If the graph is connected, we say that the persona elements (maximal frequent categorysets) extracted from the segment are compatible, and it is a candidate for yielding a persona.

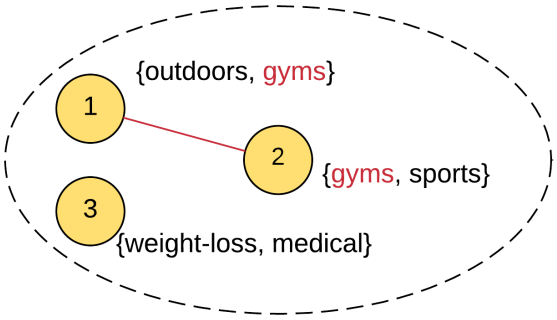

Fig. 4: An incompatible set of elements. The graph corresponding to the set is not connected.

incompatible, therefore, the corresponding segment will be flagged as not admitting any persona.

Human analysts will further filter some segments if the maximal frequent categorysets are gender or age inappropriate. For instance, if a segment has the maximal frequent categoryset $\{sexual\text{-}supplements, sexual\text{-}health\}$, it will be flagged as not having a persona since the included categories are inappropriate to be shown to general customers.

h) Step 8: Designate Personas by Human Inspection: Finally, human analysts inspect each set of compatible elements and associate a persona with these. Personas are associated using the dual criteria of descriptiveness and distinctiveness. Namely, each persona should be descriptive of its elements, but at the same time differentiate between distinct personas. This is done carefully, since we found some personas to be related, yet subtly different. In those cases, the analysts carefully understand the distinction between such sets of compatible elements before associating personas to each. Also, to ease notation, we will use $\mathsf{ELEM}(S)$ to denote a persona as well.

V. Confidence and Coherence

In this section, we associate two measures to a persona that capture the notions of applicability and coherence of a persona to a segment.

i) Measuring Confidence using Element Supports: We wish to associate a measure of confidence in the persona based on how widely it is known to be applicable in its segment. We bootstrap from the notion of confidence for frequent categorysets in order to do this. Let $p \in \mathsf{ELEM}(S)$ be a frequent categoryset.

Definition 6. The *confidence* in p is defined as $\mathrm{supp}(p)/|S|$.

Definition 7. The *confidence* in a persona $\mathsf{ELEM}(S)$ is defined as

$$\frac{\cup\{\mathrm{supp}\,(p)\colon p \in \mathsf{ELEM}(S)\}}{|S|}.$$

j) Measuring Coherence using Graph Connectivity: Let S have a persona. We associate to S a graph $R(S)$, called its *coherence graph*.

Definition 8. The vertices of $R(S)$ are the set of all categories in $\cup\mathsf{ELEM}(S)$. Two vertices are joined by an edge iff they both lie in some element $p \in \mathsf{ELEM}(S)$.

Definition 9. The *coherence* of a persona associated to a segment S is defined as the connectivity $\kappa[R(S)]$ of the graph $R(S)$. Namely, it is the size of the minimum vertex cut of $R(S)$.

The larger is $\kappa[R(S)]$, the more intertwined are its elements, and consequently the more coherent the persona is as a description of its corresponding users.

We return to the example of Fig. 3 to explain the calculation of coherence of a persona. The graph representation $R(S)$ of the maximal frequent categorysets $\{clothing, jewelry\}$, $\{jewelry, nightlife\}$, and $\{nightlife, bars\}$, is a line graph on four vertices having three edges. Any of the vertices of degree two forms a minimum vertex cut. Therefore, according to Def. 9, the coherence of this persona is one (namely, the size of the minimum vertex cut of $R(S)$).

VI. Large Scale Empirical Work

In this section, we clarify our implementation choices, and then provide examples of mined personas.

A. Implementation and Parameter Choices

User representation. We set $\alpha = 0.995$ for time decay and the $\gamma = 5$ for activity integration. These values are suggested by A/B tests in our e-commerce system. The dimension M of user representations is 140: this being the number of categories.

User selection. We exclude users having less than ten clicks and having no purchase in the past one month. This threshold is obtained by balancing the percentage of users being considered, and the average number of activities of the selected users.

In the experiments, we initially consider users who have viewed any product since July 1st, 2014, totalling around 38 million users. From these, around 16 million users are selected by applying the aforementioned filtering criterion.

User segmentation. We first length-normalize the activities vector of each user to unit length, and cluster using K-means with cosine similarity metric. The optimum number of segments is obtained using the *elbow criterion* [15]. Namely, first we plot the clustering coefficient[2] w.r.t. different values

[2]Defined as $\sum_{i,j;U_i \in S_j} \cos(U_i, \mu_j) - \frac{1}{K-1}\sum_{i,j;U_i \notin S_j} \cos(U_i, \mu_j)$, where μ_i are cluster (segment) centroids, and U_i are user vectors.

TABLE I: Examples of lifestyle personas obtained from 90 user segments (§VI-B). Personas are descriptive and distinctive.

Lifestyle Persona[a] (by Human Analysts)	Segment Size (percentage of users)	Elements (maximal frequent categorysets)	Coherence (connectivity $\kappa[R(S)]$)
APPEARANCE CONSCIOUS[b]	1.77	{clothing-and-shoes,womens-clothing} {jewelry,womens-clothing} {activities,womens-clothing}	1
HEALTH CONSCIOUS[c]	1.54	{cosmetic-procedures,medical,weight-loss}	2
SKIN CONSCIOUS[c]	1.22	{cosmetic-procedures,skin-care} {salons,skin-care}	1
HOME IMPROVER	0.85	{furniture-stores,home-and-garden-local}	1
NIGHTLIFE LOVER	0.62	{activities,beer-wine-and-spirits,nightlife} {beer-wine-and-spirits,nightlife,restaurants}	2
ACCESSORIES LOVER	0.38	{mens-accessories, womens-accessories}	1

[a] Color code: orange: dominated by local deals; green: having both local and goods components; blue: dominated by goods
[a] Local categories dominate personas. Around 65% of personas have Local categories only, and over 75% of personas have at least one Local category. In contrast, Goods categories appear in around 35% of personas.
[b] Note that though appearance conscious, they are not focused on health. Compare to HEALTH CONSCIOUS, who are also appearance conscious, but focused on health aspects such as weight [c] Distinctiveness of personas illustrated by two subtly different personas: both are health conscious, but one has a focus on skin

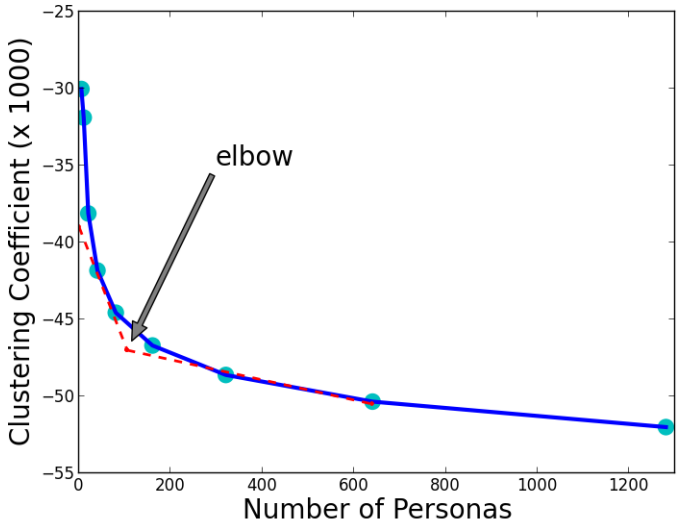

Fig. 5: Elbow criterion to determine optimal number of segments using clustering (see §VI-A).

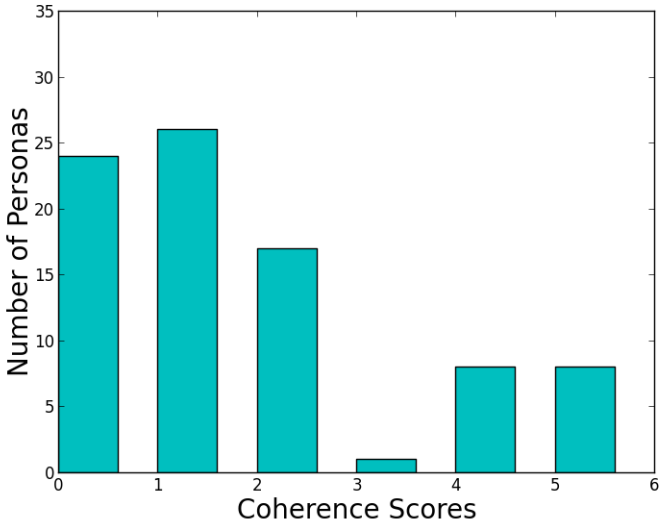

Fig. 7: The distribution of coherence in our mined personas (see §VI-A).

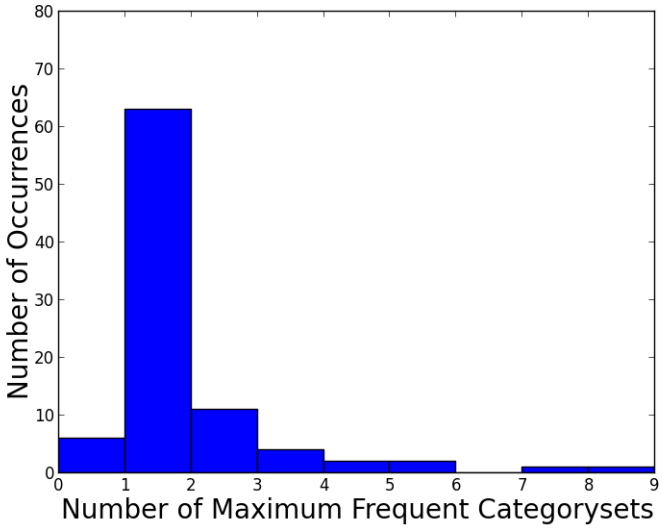

Fig. 6: Distribution of number of maximal frequent categorysets in segments that produce personas (see §VI-A).

of the number of clusters, as shown in Fig. 5. We see that the clustering coefficient decreases when the number of personas increases. When the number of personas is around 100, the marginal loss drops, and adding more clusters does not yield better modeling of the data. Experimenting further around the value of 100, we found 90 to be an optimal choice in terms of producing significant clusters. For the purposes of this paper, we used the MapReduce K-Means in Apache Mahout[3] for the sake of simplicity in sharing the design.

Element extraction. We use MapReduce Apriori [6] on each segment S in order to extract $\mathsf{ELEM}(S)$. We tested our algorithm using different values of T including 5%, 10%, 15% and 30%. We compared the output user personas. The user personas remain almost unchanged when using 5% to 15%. Therefore, we set the threshold to be 15% for the highest confidence of the algorithm.

Compability of elements. Out of our 90 clusters, one cluster produces no frequent categorysets at the thresholds we set (and thus has no persona associated). Five clusters have maximum frequent categorysets that are not compatible, thus also have no personas. The rest yield personas. Fig. 6 illustrates the distribution of maximal frequent categorysets in segments that produce personas.

Confidence and coherence. Although we are unable to share confidence distributions for our personas (these being company confidential), we can state that a significant number of personas have confidence exceeding 50%.

The histogram of the coherence scores of our personas is shown in Fig. 7. The coherence scores of our personas range from zero through five. When the score is zero, there is only one category in the persona. For most of the cases (67 out of 84 learned personas), the coherence scores are less than 3. According to Def. 9, this means that, for most of the personas, the minimum vertex cuts have less than three categories.

B. Examples of Personas

Table I provides examples to show the results of user persona mining in our e-commerce service. Discussion and copious observations are recorded as table-notes.

C. Components of Lifestyle Personas: Local vs. Goods

Comparing the components of lifestyle personas shown in Fig. 8, we notice that local categories dominate personas. Around 77% of personas have at least one local category, and around 31% of personas have at least four local categories. In contrast, goods categories appear in around 36% of personas. Local categories occur more than twice as often as goods

[3]https://mahout.apache.org/

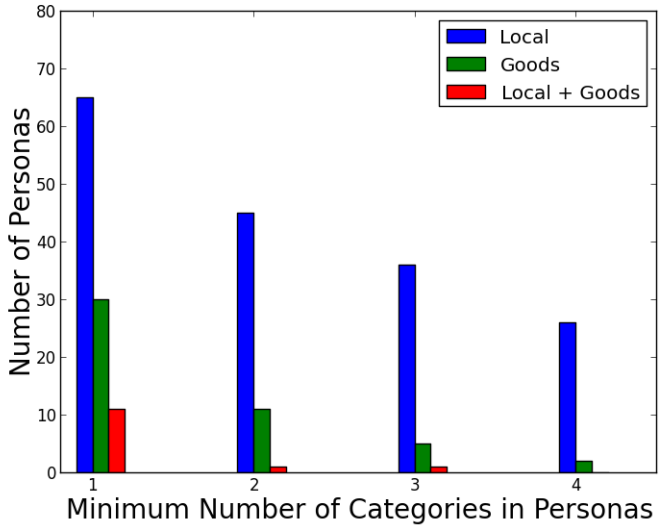

Fig. 8: The components of lifestyle personas: local vs. goods (see §VI-C).

categories, across all the cases. Note that Fig. 8 shows absolute numbers, not percentages.

D. Unusual Personas

Our user persona mining reveals some surprising combinations of leaves in the product taxonomy. Seemingly unrelated leaves of the taxonomy are sometimes bound together to form personas. As an example, *mens-shoes, alcohol, tobacco* is identified as the persona for a group of users making up about 0.2% of population. It is easily understandable that *alcohol* and *tobacco* are associated, but it was not clear from either a marketing aspect or an engineering aspect that *mens-shoes* is also associated with *alcohol* and *tobacco*. As another example, *auto-repair, vision* is the persona for a group of users making up roughly 0.9% of population. It was also hardly noticed that *auto-repair* and *vision* are strongly related to each other for a significant group of customers.

VII. Conclusion

We introduce the problem of mining "lifestyle personas" from user activity data at a major e-commerce company. We present one of the largest-scale studies on industrial mining for segmentation and related tasks.

The mined personas are descriptive and distinctive, and help advertisers, publishers, and online sellers to understand the compositions and characteristics of their customers. The machinery of personas (including combinations of confidence and coherence measures for these) is being productionized at Groupon to power a range of applications from collaborative filtering based recommendations to designing the UI for new users. We believe this is a standout example of the application of data mining to real-world industrial problems at large scales.

References

[1] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[2] M. Chen, A. Chiu, and H. Chang. Mining changes in customer behavior in retail marketing. *Expert Systems and Applications*, 28(4):773–781, 2005.

[3] R. Gowri, A. J., A. Kumar, and J. R. K. An overview of clustering algorithm and collaborative filtering method through e-commerce data perspective. *Proc. IJERT*, 4(1):75–79, 2015.

[4] J. Grudin and J. Pruitt. Personas, participatory design, and product development: An infrastructure for engagement. *Proc. PDC*, pages 1–15, 2002.

[5] J. Hair, R. Anderson, R. Tatham, and W. Black. *Multivariate Data Analysis: A Global Perspective*. Prentice Hall, 2009.

[6] N. Li, L. Zeng, Q. He, and Z. Shi. Parallel implementation of apriori algorithm based on mapreduce. *Proc. SNPD*, pages 236–241, 2012.

[7] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

[8] W. Liu, F. Jin, and X. Zhang. Ontology-based user modeling for e-commerce system. *Proc. PERVASIVE*, 1:260–263, 2008.

[9] A. Pereira and E. Hruschka. Simultaneous co-clustering and learning to address the cold start problem in recommender systems. *Knowl.-Based Sys.*, 82:11–19, 2015.

[10] J. Pruitt and J. Grudin. Personas, practice and theory. *Proc. DUX*, pages 1–13, 2003.

[11] A. Savvopoulos, M. Virvou, D. Sotiropoulos, and G. Tsihrintzis. Clustering for user modeling in recommender e-commerce application: A rup-based intelligent software life-cycle. *Proc. JCKBSE*, pages 295–304, 2008.

[12] H. Stormer. Improving e-commerce recommender systems by the identification of seasonal products. *Proc. AAAI*, pages 92–99, 2007.

[13] Q. Su and L. Chen. A method for discovering clusters of e-commerce interest patterns using click-stream data. *Electronic Commerce Research and Applications*, 14(1):1–13, 2015.

[14] V. Thoma and B. Williams. Developing and validating personas in e-commerce: A heuristic approach. *Proc. INTERACT*, 5727:524–527, 2009.

[15] R. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[16] R. Wu and P. Chou. Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications*, 10(3):331–341, 2011.

[17] Z. Xizheng. Building personalized recommendation system in e-commerce using association rule-based mining and classification. *Proc. ICMLC*, 7:4113–4118, 2007.